Identification and Analysis of Cell Cycle Phase Genes by Clustering in Correspondence Subspaces*

Ai Sasho[1], Shenhaochen Zhu[2], and Rahul Singh[2,**]

[1] Department of Chemistry and Biochemistry
[2] Department of Computer Science
San Francisco State University, San Francisco, CA
rahul@sfsu.edu

Abstract. Correspondence analysis (CA) is a statistical method that is widely used in multiple disciplines to reveal relationships amongst variables. Among others, CA has been successfully applied for microarray data analysis. One of CA's strengths is its ability to help visualize the complex relationships that may be present in the data. In this sense, CA is a powerful exploratory tool that takes advantage of human pattern analysis abilities. The power of CA can, however, be diluted, if the patterns are embedded in data clutter. This is because CA is a dimensionality reduction approach and not a data reduction method; thus, is powerless to remove clutter. Unfortunately, our visual analysis abilities can be overwhelmed in such conditions causing failures in identifying relationships. In this paper, we propose a solution to this problem by combining CA with one-way analysis of variance (ANOVA) and subsequently by clustering in the low-dimensional space obtained from CA. We investigate the proposed approach using microarray data from 6200 *S. cerevisiae* genes and demonstrate how visual analysis is facilitated by removal of unnecessary clutter as well as facilitating the discernment of complex relationships that may be missed through application of CA alone.

Keywords: Correspondence Analysis, UPGMA Clustering, One-way ANOVA, Microarray Time Course Data.

1 Introduction

Microarray technology enables the analysis of the mRNA levels of thousands of genes simultaneously providing a powerful tool for researchers. This technology has been widely used and became a standard tool for studying the fundamental aspects of growth and development of organisms as well as the genetic causes of many diseases. Microarrays provide an opportunity to study interactions not only among genes but also relationships between genes and experimental conditions. In microarray experiments, typically, gene expressions of thousands of genes are measured over a period of time, accumulating a large volume of data very quickly. Thus, such

[*] This research was funded in part by the NSF grant IIS-064418 (CAREER).
[**] Corresponding author.

A. Abraham et al. (Eds.): ACC 2011, Part I, CCIS 190, pp. 340–350, 2011.

experiments necessitate the development of efficient and automated way of analyzing the data. Furthermore, these analysis techniques must also ultimately help in data interpretation and data exploration; revealing the issues that need to be explored further. Consequently, solutions need not only to address issues of automation and algorithmic efficacy, but must also aid the human-algorithm interface.

In this paper, we propose a method to address the somewhat dichotomous requirements of the above design formulation. In doing so, we base ourselves on a dimensionality reduction technique called correspondence analysis (CA) which has also been shown to be of promise in microarray data analysis [2]. CA offers a way of revealing the relationship between and among variables in a data set by applying a specific form of dimensionality reduction and produces a graph including all the variables in a single low-dimensional subspace. For instance, using CA, one may simultaneously visualize relationships between genes and hybridizations as well as within genes and within hybridizations. However, the interpretation of the visual results generated by CA still requires human intervention. The human visual system, though extremely powerful in discerning patterns, is not very efficacious in environments containing a large amount of self-similar data (cluttered environments). To address this issue, we propose the use of ANOVA and clustering in the reduced-dimension space identified using CA. Our goal is to accentuate the meaningful data (signal) while minimizing the background data clutter and thus reduce the cognitive load during analysis. We begin this paper in the following with a review of related work and distinctions of the proposed approach from them. This is followed by a detailed description of the proposed method and experimental analysis of its performance.

1.1 Prior Research and Overview of Proposed Method

The microarray data analysis is one of the most heavily research areas in contemporary bioinformatics. However, most methods that have been proposed for this problem can be thought of as belonging to one of the following classes [2]: (1) clustering methods, (2) dimensionality reduction techniques, and (3) techniques that treat the problem as that within the classification/regression framework. Of these, the dimensionality reduction methods are of direct relevance for us. The commonly used methods in this class include principle component analysis (PCA) and Multi-dimensional scaling (MDS). PCA utilizes the properties of covariance matrix and transforms correlated variables into orthogonal (uncorrelated) variables while preserving as much information as possible present in the original data set. The use of PCA in microarray data analysis was demonstrated by Raychaudhuri *et al.* [5] who successfully applied PCA to a sporulation time series microarray data to find the temporal gene expression patterns. MDS is a set of statistical methods to reveal the underlying structure of the data set by using a dimensionality reduction technique. MDS has been used in many applications of data mining; however, the computational complexity of MDS makes it difficult to apply the method to a large set of data [8]. In this context, Tzeng *et al.* [8] has developed a modified MDS which reduces the computational complexity of MDS, and proved effectiveness of MDS to expose correlation of certain human genes.

Like these two methods, CA is also a technique that belongs to the class of dimensionality reduction approaches. However, unlike the above class of methods, CA has certain properties that make it more suitable for revealing association among variables. Specifically, CA can aid in investigating the association amongst variables (row and column variables) by projecting them in single joint space. Furthermore, CA has a low computational complexity. Thus, if our goal is to study the interdependencies between genes and hybridizations (experimental conditions), then correspondence analysis is arguably a very apt technique. This conclusion was demonstrated by Fellenberg *et al.* in their seminal paper [2]. Fellenberg *et al.* applied CA to the *Saccharomyces cerevisiae* gene expression microarray data produced by Spellman *et al.* [6], and successfully showed that CA can be used to visualize relationships between genes and experimental conditions. However, the method in [2] did not address the issue of data clutter. Furthermore, dimensionality reduction approaches (CA included) require a complete data matrix and cannot function in the presence of missing information. However, it is inevitable to avoid missing values when dealing with microarray data. Such questions were also not considered in [2]. Finally, at the current state-of-the-art, the question of algorithmic analysis beyond application of CA has not been considered. In our work, we address all these questions in context of the problem of relating the genes of *S. cerevisiae* to the specific cell-cycle phases in which they are expressed. For this purpose, the microarray data compiled by Spellman [6] is represented in a matrix containing genes in rows and cell cycle phase time points in columns.

In the proposed method, first, the missing data problem is addressed. To replace missing values, we investigate several missing value estimation techniques, including row average, cell cycle row average, and Bayesian Principle Component Analysis (BPCA) (http://hawaii.sys.i.kyoto-u.ac.jp/~oba/tools/BPCAFill.html). These methods are evaluated in terms of the percentage of correctly associated gene-cell cycle phase pairs after correspondence analysis. Thus, our assessment studies the impact of these techniques on the actual analysis. Our assessment criterion is different (and arguably richer) than the standard approach of synthetically removing data and using linear error measures (such as RMSD) to judge efficacy. We next address the issue of ameliorating data clutter in two steps; the first step occurs before application of CA. In this step, we apply ANOVA to the microarray data to identify the genes showing differential expression. The effectiveness of ANOVA in determining genes with differential expression was proved by Cui *et al.* [1]. By performing ANOVA, we reduce the data (by filtering out genes that are expressed in the constant levels) without negatively influencing subsequent analysis. Furthermore, removal of non-differentially expressed genes also reduces the computational requirements from subsequent analysis steps and simplifies visual analysis. Next, we perform the dimensionality reduction step by using CA. This specific step is similar to the work in [2]. Following CA, we perform clustering in the low dimensional space to further reduce data clutter and aid in interpretation. In this paper, we use UPGMA algorithm to further study relationships between genes. We note that other clustering algorithms are equally applicable. Our choice of the specific clustering method is motivated by two reasons: first, the UPGMA algorithm is well understood leading to easier analysis of the final outcomes, especially as the clustering is performed on a reduced dimensional yet information-rich subspace. Second, the input data for UPGMA is a distance matrix which can be easily constructed by calculating the distances between all data points in the correspondence subspace. The UPGMA produces an ultrametric tree, depicting the relationships among data elements. Analysis of the tree can provide additional information about the associations among variables which cannot be discerned easily from the output of correspondence analysis alone.

2 Proposed Approach

2.1 Data Set

To evaluate our implementation of computational methods, we use the microarray data of *Saccharomyces cerevisiae* collected by Spellman *et al.* [6] as an input data. The data set contains the color intensities of approximately 6200 *S. cerevisiae* genes whose gene expression was synchronized by four different synchronization methods: α-factor, *CDC15*, *CDC28* and elutriation. For each method, the gene expression was recorded every 10 minutes for up to 390 minutes. Each time point (10 minute interval) is mapped to the biological cell cycle phases, namely M/G1, G1, S, S/G2 and G2/M [6]. A CSV file containing the microarray data is created and used as an input file for the program developed in this paper.

2.2 Missing Value Estimation

Correspondence analysis requires a complete set of data. Unfortunately, gene expressions measured by microarrays often include missing values; thus, a missing value estimation step is necessary for further analysis. Three missing value estimation methods are evaluated as part of our investigations: (1) missing value estimation by row average, (2) by cell cycle row average, and (3) by the Bayesian Principle Component Analysis Missing Value Estimator [4]. The effectiveness of a missing value estimation method is evaluated by computing the percentage of correctly associated gene-cell cycle phase pairs based on the microarray data published by Spellman et al. [6]. For the missing value estimation by row average, the average gene expression value of each row is calculated and used to fill in the missing values. Similarly, missing value estimation by cell cycle row average imputes the missing values by calculating the average of the row by only using the columns that belong to the same phase as the missing value. The BPCA Missing Value Estimator is a publicly available program developed by Oda *et al.* [4].

2.3 Identification of Non-differential Genes Using ANOVA

Our interest is to identify genes showing different temporal profiles through cell cycle phases and associate these genes with a specific cell cycle phase. Therefore, genes that are constantly expressed at the same level throughout the cell cycle phases, such as house keeping genes, can be filtered out prior to correspondence analysis and UPGMA. After estimating missing values, the one-way ANOVA method is applied to the microarray data to identify the genes that are showing differential expression. The *F* value calculated by ANOVA is evaluated against the critical *F* value to determine if genes are expressed differently through cell cycle phases.

The microarray data contains genes in rows and cell cycle phase time points in columns. Since the columns represent cell cycle phase time points, they can be categorized by cell cycle phases, and the columns belonging to the same cell cycle phase can be considered as a group in the ANOVA process. To perform ANOVA, the total sum of square is calculated. The total sum of square is defined by the following equation:

$$SS_{total} = \sum X^2 - \frac{G^2}{N} \qquad (1)$$

In the above equation, X represents the sum of squared data points in a group; G is the sum of all the data points; and N is the total number of data points. Next, the squared sum within a group is calculated in Eq. (2), where n is the number of data points in a group. Similarly, the squared sum between the groups is calculated in Eq. (3) below.

$$SS_{within} = \sum X^2 - \frac{(\sum X)^2}{n} \qquad (2) \qquad\qquad SS_{between} = \sum \frac{T^2}{n} - \frac{G^2}{N} \qquad (3)$$

where T is the sum of data points for each group. After calculating the above values, SS_{total} should equal the sum of $SS_{between}$ and SS_{within}. The means of squares within and between are calculated by:

$$MS_{within} = SS_{within} / df_{within} \qquad (4) \qquad\qquad MS_{between} = SS_{between} / df_{between} \qquad (5)$$

where df stands for the degree of freedom (which by $N - 1$). By using the means of squares within and between, F-statistics value is calculated as follows:

$$F = \frac{MS_{between}}{MS_{within}} \qquad (6)$$

The F-statistics value indicates the variance among the groups. Therefore, genes with a higher F-statistics value than the critical F value show differential expression during the cell cycle and are subject to further analysis by correspondence analysis.

2.4 Correspondence Analysis

CA is applied to the genes that are identified by the ANOVA process as genes showing differential expression. Here, by using correspondence analysis, our aim is to associate a gene with a specific cell cycle phase by identifying a cell cycle phase in which the gene is up-regulated or down-regulated. The microarray data is represented as I x J matrix containing genes in rows and cell cycle phase time points in columns. The symbols I and J denote for the number of genes and the number of cell cycle phase data points respectively. A datum in a matrix at the ith row and the jth column is written as n_{ij}. After a series of matrix manipulations, correspondence analysis calculates coordinates of the row variables (genes) and the column variables (cell cycle phases), which are then used to plot a graph in the desired dimension [3]. The main steps of correspondence analysis are the followings [2, 3]:

Step 1: The mass of each column and row are calculated. The mass of a column is defined as the sum of the data elements in the column divided by the sum of all data elements.

$$c_j = n_{+\mathrm{j}} / n_{++} \quad (7) \qquad r_i = n_{i+} / n_{++} \quad (8)$$

In Eq. (7), c_j is the mass of the *j*th column, $n_{+\mathrm{j}}$ is the sum of the data at the *j*th column, and n_{++} denotes the sum of all the elements in the matrix. Similarly, the mass of a row is calculated in Eq. (8), where r_i represents the mass of the *i*th row, and n_{i+} is the sum of the data in the *i*th row.

Step 2: A correspondence matrix *P* is calculated by dividing each datum by the sum of all the data elements as shown in Eq. (9).

$$p_{ij} = n_{ij} / n_{++} \quad (9) \qquad s_{ij} = (p_{ij} - r_i c_j)\sqrt{r_i c_j} \quad (10)$$

In Eq. (9), p_{ij} is a value in the correspondence matrix at the *i*th row and the *j*th column, and n_{ij} represents a data point at the *i*th row and the *j*th column in the original matrix.

Step 3: By using the values form the correspondence matrix, the singular matrix *S* is derived using Eq. (10), where s_{ij} represents a value in the singular matrix and r_i and c_j are the mass of the *i*th row and the *j*th column respectively.

Step 4: The matrix *S* is factored using singular value decomposition (SVD). We use the Java Matrix package (http://math.nist.gov/javanumerics) to compute the SVD. As a consequence of the SVD, the matrix *S* is decomposed into three matrices *U*, *Λ*, and *V* as shown in Eq. (11).

$$S = U\Lambda V^T \quad (11)$$

In Eq. (11), *U* denotes the matrix containing left singular vectors, *Λ* stands for a diagonal matrix containing diagonal elements in a sorted order, and *V* denotes the matrix containing the right singular vectors.

Step 5: The values from the *U*, *Λ*, and *V* matrices are used to determine a 2D mapping of the data where the row variables (genes) are mapped to the x-axis and the column variables (cell cycle phases) are mapped to the y-axis. The row variable (gene) coordinates are calculated as shown in Eq. (12).

$$f_{ik} = \lambda_k * \upsilon_{ik} * \sqrt{r_i} \quad (12) \qquad g_{jk} = \lambda_k * v_{jk} * \sqrt{c_j} \quad (13) \qquad Inertia = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \quad (14)$$

In Eq. (12), f_{ik} is a gene coordinate at the *i*th row and the *k*th column where $k = 1,\ldots, J$. The λ_k denotes a diagonal element in the singular matrix *Λ* at the *k*th position. The υ_{ik} represents a value in the *U* matrix at the *i*th row and the *k*th column. Similarly, the coordinates of the column variables (cell cycle phases) are calculated using Eq. (13), where g_{jk} is a cell cycle phase coordinate at the *j*th row and the *k*th column with $k = 1,\ldots, I$, and v_{jk} stands for a value in the *V* matrix at the *k*th position. The gene coordinate and cell cycle phase coordinate matrices are multidimensional matrices, and each column represents a dimension. For instance, the first column contains the x-axis values (the 1st dimension), and similarly the second column contains the y-axis values (the 2nd dimension). Since a two-dimensional graph is plotted in this paper, the values from the first two columns are used for drawing a biplot. This constitutes a dimensionality reduction step. Although correspondence analysis retains information present in the original data as much as possible, some information is lost during the dimensionality reduction process. The information lost in the process is called inertia, and is calculated by Eq. (14).

2.5 UPGMA Clustering in the Correspondence Subspace

UPGMA is a hierarchical clustering algorithm for grouping data points based on distances between elements in a cluster. UPGMA takes a dissimilarity matrix as an input and joins the nearest clusters until only one cluster is left. First, the nearest clusters are identified by finding the pairwise minimum distance between elements. These clusters are removed from the dissimilarity matrix, and a new joint cluster is inserted. The new cluster contains the average distances between elements in two clusters that are merged. The procedure to join nearest clusters and compute distances for the new cluster is repeated until all the clusters are joined. The information about joined clusters and the minimum distances are saved through the iterations and used to construct an ultrametric tree to show the relationships between the elements.

3 Experimental Investigations and Results

The proposed method was applied to the yeast gene microarray data [6] to demonstrate the effectiveness of the combined methods. The data set contains 6179 genes in rows and 73 cell cycle phase time points in columns, and the columns are categorized into five different cell cycle phases. We evaluated our results by comparing our gene-cell cycle associations against the list of gene-cell cycle pairs created by Spellman *et al.* [6]. The percentage of correctly associated gene-cell cycle pairs was calculated using the metrics of *precision* and *recall*. Precision is the fraction of correctly associated gene-cell cycle pairs within the data set analyzed by the program. On the other hand, recall refers to the fraction of correctly associated gene-cell cycle pairs in the data set containing all the known gene-cell cycle pairs identified by Spellman *et al.* [6]. In this section, we present the results obtained from missing value estimation, ANOVA, correspondence analysis and UPGMA, following the order the procedures were applied.

3.1 Missing Value Estimation

Three missing value estimation techniques: estimation by row average, cell cycle row average and BPCA Missing Value Estimator [4], were evaluated in this project. Each technique was assessed by calculating the percentage of genes that were associated to a correct phase after the CA step. By doing so, our goal was to assess the impact of the missing value estimation method on the overall analysis. The results are shown in Fig. 1.

Missing value estimation by cell cycle row average produced approximately 77.50% and 69.84% of correct gene-cell cycle phase pairs in precision and recall respectively, producing the best results among the techniques evaluated in this paper. The row average method produced about 77.44% and 69.59% in terms of precision and recall, and BPCA Estimator resulted in the precision of 76.59% and the recall of 69.21%.

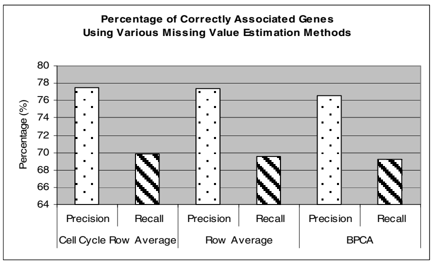

Fig. 1. The graph shows the precision and recall in percentage using different missing value estimation techniques after CA

3.2 Influence of ANOVA

In the ANOVA process, 3054 genes were identified as genes showing differential expression using the *F* critical value at $p = 0.3$. This included approximately 90.11% of the differentially expressed genes identified by Spellman *et al.* [6]. These genes constituted our “genes of interest” [7] and served as an input data set for correspondence analysis. To evaluate the effectiveness of ANOVA, the percentages of correctly associated gene-cell cycle phase pairs were compared with and without ANOVA. Without using ANOVA, the percentage of genes assigned to a correct phase after CA was approximately 28.79%, and the ratio was increased to 69.84% when ANOVA was incorporated.

3.3 Analyzing the Data Using Correspondence Analysis and Clustering

The biplot in Fig. 2 was produced by applying correspondence analysis on the genes of interest. In the graph, genes are represented by black dots, and the cell cycle phase data points are in various shapes according to the assigned phase. The cell cycle phase centroids were calculated by plotting the average x-coordinate and y-coordinate of the data points belonging to each cell cycle phase. The lines were extended from the origin of the graph to the centroids, so that the user can readily recognize the scattering pattern of cell cycle phase data points. For purposes of the comparison with Fellenberg *et al.* [2], we added the labels to the genes that are known to participate in histone production. Histones are used to coil strands of DNA; thus, histone related genes should be up-regulated during the DNA synthesis phase, as shown in Fig. 1. There is an empty spot surrounding the origin of the graph. In a correspondence analysis graph, data points closer to the origin of the graph do not show a strong association to any of the other data points. In the context of our problem, the genes near the center do not show differential expression. The ANOVA step removed such genes prior to the application of CA. It should be noted that from a visual analysis perspective, this reduces unnecessary clutters due to genes that are irrelevant (non-differentially expressed). Our CA biplot resembled the biplot produced by Fellenberg *et al.* [2]. A visual inspection of two biplots reveled that the locations of cell cycle phase data points, histone genes, and the order of cell cycle phase clusters show similarities between two graphs.

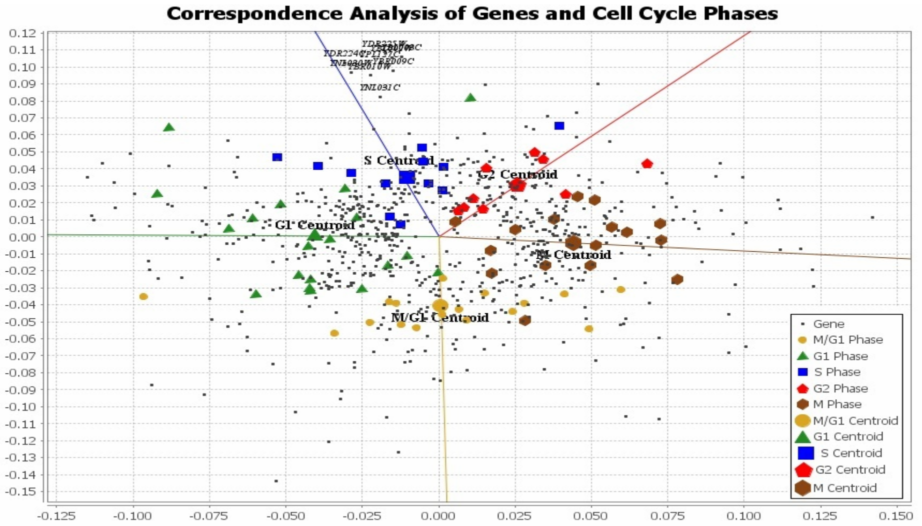

Fig. 2. Biplot produced by correspondence analysis. Black dots represent genes and the symbols represent cell cycle phase time points. Lines were extended to centroids. The histone genes cluster around the S cell cycle phase.

In CA, genes were associated to a cell cycle phase by identifying the closest phase centroid and the percentage of correct gene-phase associations was calculated in terms of precision and recall. The precision measures the accuracy of our analysis given our data set, and the recall measures the accuracy against the complete data set. After application of CA, the precision was calculated to be 77.50%, and the recall was found to be 69.84%. The reader may note that in the work by Fellenberg *et al.* [2], this type of assessment of accuracy was not performed.

In the final step of our method, the data in the correspondence subspace were clustered by using the UPGMA algorithm. The coordinates of data elements in the CA biplot were used to construct a distance matrix, containing the all-pair Euclidean distances (computed in the correspondence subspace). The UPGMA algorithm was applied using the distance matrix and the tree structure in Newick format was produced by UPGMA. Fig. 3 represents a dendrogram constructed based on the Newick string. By inspecting the dendrogram, it can be noticed that the genes assigned to the same cell cycle phase cluster together, and these clusters are placed in a pattern. The most noticeable cluster is the G1 cluster (in black) occupying the large part of the second row in Fig. 3. This observation consists with the correspondence biplot produced in the CA step, which also shows a large number of G1 genes crowded together in one area. In the Fig. 3 dendrogram, the clusters appear in the order of: M (blown) (mixed with some G2), G1 (black), S (blue), G2 (red), S, G1 (the long stretch), M/G1 (orange), and M. Barring a few exceptions, this order resembles the reverse order of the cell cycle phases: M, G2, S, G1, and M/G1. We can also notice large clusters of M genes present both at the beginning and the end of the dendrogram. This is due to the circular distribution of the data in the reduced dimensional correspondence subspace as can be seen in Fig. 2.

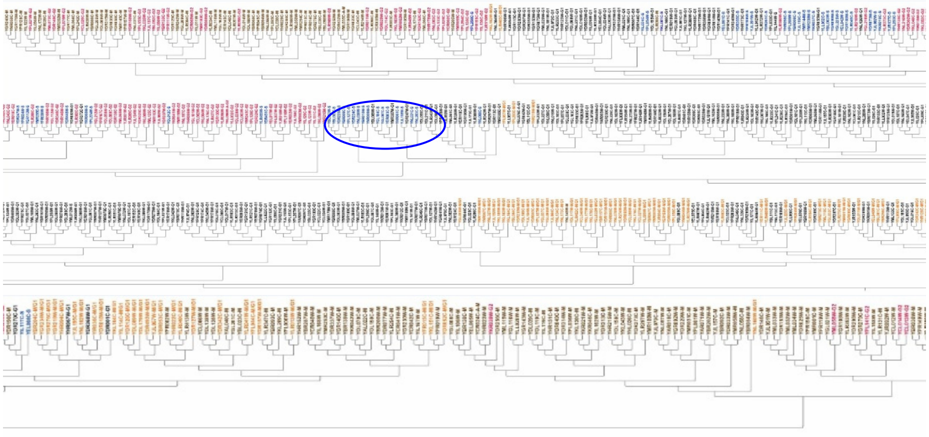

Fig. 3. The dendrogram constructed by UPGMA clustering. A leaf represents a gene. The genes are color coded according to the cell cycle phase assigned by Spellman *et al*. [6]. The color codes are: M/G1 genes in orange, G1 genes in black, S genes in blue, G2 genes in red, and M genes in brown. Histone genes are circled in blue. Note that the continuous dendrogram was divided in sections to fit in the paper.

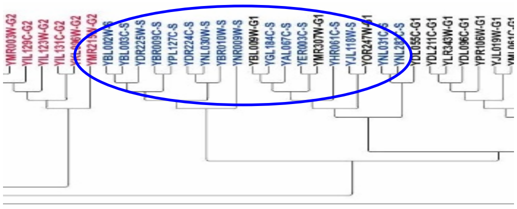

Fig. 4. The blue circle in Fig. 3 is magnified here to show a cluster containing histone genes

In the relation to the correspondence biplot, the clusters appearing in the correspondence subspace stay as clusters in the dendrogram. For example, the histone genes assigned to the S cell cycle phase formed a cluster in both CA biplot (Fig. 2) and in the dendrogram (Fig. 4). In addition, the genes surrounding the cluster centroids in the CA biplot tend to form large clusters in the dendrogram. There are some genes that are spread in a different phase cluster, e.g. a dozen G2 genes are interspersed among the M phase genes at the beginning of the dendrogram, and several M/G1 genes are spread among the long stretch of the G1 cluster. It is usually found that when random genes disturb a cluster, these random genes belong to a neighboring cluster.

4 Conclusions and Discussions

This paper demonstrates a novel approach to reveal hidden relationships among variables by combining CA with ANOVA and UPGMA-based clustering. The method provides a simple visual representation of the complex relationships in the data. Through the application of ANOVA, the accuracy of the analysis was increased by approximately 41.0% and the data set was reduced by 50.0%. In the process of applying ANOVA, about 9.0% of relevant genes were lost. CA associated approximately 77.5% of the genes to the correct cell cycle phase and produced a graph exposing associations among the data elements. The UPGMA algorithm, which was applied to the correspondence subspace, revealed additional associations not only between genes and cell cycle phases, but also within genes. UPGMA produced a hierarchical graph revealing the clusters of genes, while showing how strongly the clusters were related. One of the thrust areas of our further research is to reduce the percentage of relevant genes that were eliminated by ANOVA. The source code and input files for this project are publicly accessible at http://tintin.sfsu.edu/projects/mace.html.

Author Contributions

Research conceptualization and overall method design (RS). Correspondence analysis (AS) and ANOVA (SZ). The experiments were conducted by AS and SZ. The paper was written by RS and AS with contributions from SZ.

References

1. Cui, X.Q., Churchill, G.A.: Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 4, article 210 (2003)
2. Fellenberg, K., Hauser, N., Brors, B., Neutzner, A., Hoheisel, J., Vingron, M.: Correspondence Analysis Applied to Microarray Data. Proc. Nat. Acad. Sci. 98, 10781–10786 (2001)
3. Greenacre, M.: Correspondence Analysis in Practice. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton (2007)
4. Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., Ishii, S.: A Bayesian Missing Value Estimation Method For Gene Expression Profile Data. Bioinformatics 19(16), 2088–2096 (2003)
5. Raychaudhuri, S., Stuart, J.M., Altman, R.B.: Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. In: Pacific Symposium on Biocomputing, pp. 455–466 (2000)
6. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell 9, 3273–3297 (1998)
7. Tai, C., Speed, C.: Statistical Analysis of Microarray Time Course Data, Walter and Eliza Hall Institute of Medical Research. In: DNA Microarrays, vol. ch. 20, Taylor and Francis, New York (2005)
8. Tzeng, J., Lu, H.-S., Li, W.-H.: Multidimensional Scaling For Large Genomic Data Sets. BMC Bioinformatics 9 (2008)